Pattern Discovery for Locating Motifs in Multivariate, Real-valued Time-series Data

David Minnen, Thad Starner, Irfan Essa, Charles Isbell

Georgia Institute of Technology,
College of Computing / Interactive Computing / RIM Center
Atlanta, GA 30332, USA
{dminn|thad|irfan|isbell}@cc.gatech.edu

The problem of locating motifs in multivariate, real-valued time series data concerns the discovery of sets of recurring patterns embedded in the time series. Each set is composed of several non-overlapping subsequences and constitutes a motif because all of the subsequences are similar. This task is a natural extension of univariate motif discovery in both the symbolic and real-valued domains as previously addressed in the bioinformatics and data-mining research.

Motif discovery bears a strong resemblance to the problem of clustering time series. The key difference is that in motif discovery, the length and position of the subsequences that compose the motifs are initially unknown, whereas in the clustering problem, the sequences are given in isolation. Even if we adopt the simplifying assumption that the length of the motif members is known, thereby allowing us to extract all of the subsequences of the appropriate length, clustering the resulting set of subsequences does not yield valid motifs. First, it is not generally the case that all of the subsequences will belong to a motif. Instead, the true motifs may be sparsely distributed in the original data. Second, and more importantly, the clustering will fail because of the problem of trivial matches (Keogh 2003).

Trivial matches arise because of the overlap between successive subsequences. For instance, consider two subsequences, S_t and S_{t+1}, of length L. These subsequences are likely to be very similar simply because they have so many data points in common (specifically, they share $L-1$ frames). If we extract all subsequences of length L, typical clustering algorithms will not lead to meaningful clusters because they do not account for the trivial matches. Instead, we should only cluster the correct, non-overlapping subsequences, but these subsequences are initially unknown. Also note that trivial matches are not a significant problem for motif discovery in symbolic data. The problem arises in real-valued time series because such data is typically smoothly varying through time. In the symbolic case, a small offset typically leads to a drastic change (*i.e.*, low similarity) because the symbols are not strongly correlated along the sequence.

Previous algorithms have addressed this problem by explicitly searching for similar pairs of non-overlapping subsequences. Motif occurrences are then located by searching for other non-overlapping subsequences within a fixed distance (called the neighborhood size) from the initial pair (Chiu 2003, Minnen 2007). Another proposed approach discretizes the time series and then searches for recurring patterns in the resulting symbolic sequences (Minnen 2006). A third approach developed in the bioinformatics literature frames the problem of motif discovery as one of greedy mixture learning. An initial mixture component provides a rough model of all of the subsequences, and then subsequent components model the actual motifs (Blekas 2003).

The difficulty with the first approach is in specifying (or estimating) the neighborhood size, while the discretization used by the second approach is computationally expensive and introduces approximation errors that can mask some motifs. Greedy mixture learning provides a more principled formulation of motif discovery but that work did not deal with the trivial match problem since it was developed for symbolic sequences.

In this work, we formulate a unifying view of motif discovery as a problem of temporal density estimation. Each motif constitutes a set of similar subsequences within the time series data. If we consider the space of all subsequences, a motif can be seen as an area of high density. The precise

meaning of "high density" can be based on an estimation of the global background density, based on estimation of the local density surrounding the density peak, determined by task-specific constraints, or determined interactively as in an active learning framework.

Viewed as a problem of mode finding according to the density of subsequences, we can reinterpret the existing methods described above. The similarity based methods utilize a rough approximation for locating density modes by substituting a search for the most similar pairs of subsequences. The high density region is then assumed to be spherical as implied by a single neighborhood size for each motif. By discretizing the time series, Minnen et al. (2006) locate modes by finding sets of similar subsequences using a very simple approximation to the underlying similarity measure. A more complex model for each high density region is used in that work, however, since hidden Markov models (HMMs) are learned from the resulting sets of subsequences.

Finally, the greedy mixture learning approach can be seen as estimating the background density with the first component and then searching for high density regions corresponding to additional components that dramatically increase the total data likelihood given the full mixture model. Our approach uses the interpretation of motif discovery as temporal density estimation to formulate a more accurate algorithm suitable for use with multivariate, real-valued time series data. Density is estimated by computing the distance to the *k* nearest, non-overlapping neighbors of each subsequence. Local maxima in the estimated density are used as candidate motifs and a HMM is estimated from the subsequences in the surrounding high density region. The HMMs then become components in a temporal mixture model, augmented by a single background model learned from the entire data set. The temporal mixture model is then fit to the time series data using established methods from the continuous speech recognition community.

Our approach combines efficient, explicit search for high density regions to initialize motif models and then utilizes a background model and competitive data explanation to ensure that the discovered motifs are distinct from the background noise and other, non-recurring subsequences. We are currently running experiments to evaluate our approach on time series data collected in a variety of domains including human activities captured by an on-body sensor, speech, and American Sign Language recorded by a video camera.

References:

Blekas, K., Fotiadis, D., and Likas, A (2003). Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, 19: 607-617, 2003.

Chiu, B., Keogh, E. & Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs. In the *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. August 24 - 27. Washington, DC, USA. pp 493-498.

Denton, A. (2004). Density-Based Clustering of Time Series Subsequences. In *3rd Workshop on Mining Temporal and Sequential Data (TDM)*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA. Aug 22.

Keogh, E., Lin, J., Truppel, W. (2003) Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research. *Int. Conf. on Data Mining (ICDM03)*, p115-122, 2003.

Minnen, D., Starner, T., Essa, I., and Isbell, C. (2006) Discovering Characteristic Actions from On-Body Sensor Data. *Int. Symp. on Wearable Computing (ISWC06)*, Montreux, CH, October 11-14, 2006.

Minnen, D., Starner, T., Essa, I., and Isbell, C. (2007) Improving Activity Discovery with Automatic Neighborhood Estimation. *Int. Joint Conf. on Artificial Intelligence (IJCAI07)*, Hyderabad, India, January 6-12, 2007.

Category: **Learning algorithms**

Presentation: **Poster/Oral.**